DOCUMENT RESUME

ED 182 617 CG 014 102

AUTHOR Holland, John L.
TITLE The Influence of Vocational Interest Inventories: Some Implications for Psychological Testing.
PUB DATE 4 Sep 79
NOTE 15p.: Paper presented at the Annual Convention of the American Psychological Association (87th, New York, NY, September 1-5, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bias: *Interest Tests: Occupational Tests: *Psychological Tests: *Psychologists: Public Opinion: Sex Discrimination: State of the Art Reviews: *Testing: *Vocational Interests

ABSTRACT

The impact of psychological testing on the public is explored. Topics discussed include a review of the controversy over interest inventories, what has been learned from the research and from the political activity that was enacted relative to employment/admission testing, and some ideas for increasing the ability to cope more readily with current controversies over tests and inventories. Conclusions indicate a need for psychologists to concentrate not only on inventory reliability and construct validity but also on the improvement of their public image. (Author/BMW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Influence of Vocational Interest Inventories: Some Implications for Psychological Testing

John L. Holland

Johns Hopkins University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

John L. Holland

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED182617

I had a lot of difficulty in preparing this talk. Frankly, I have always had a lot of difficulty in preparing talks. In this instance, the difficulties have been especially frustrating. The topics for these centennial symposia call for macro-thinking and wisdom. I am not even a good micro-thinker, despite many years of 14-page, 3-table efforts.

I was asked to stimulate discussion with the audience. I believe my talk will do that, but I am not prepared to respond in a knowledgeable manner to every issue that I will raise. The area of psychological testing appears now to be too broad and too specialized for a single person to comprehend it all. If you would like to clarify my discussion or argue for another view, I will be happy to listen and learn.

The chief goal of my talk is to illuminate the impact of psychological testing on the public. I hope to achieve this goal by a review of the interest inventory controversy that began in 1972, peaked in 1975, and died in 1978. That controversy contains all the basic elements contained in the I.Q., the admissions testing, and the employment testing controversies. In short, I will begin with the interest inventory controversy. Then I will summarize what we have learned from the research and political activity that was part and parcel of this social-psychometric brouhaha. And last, I will conclude with some ideas and recommendations for increasing our ability to cope more expeditiously with the current controversies about tests and inventories. I believe these suggestions can lead to better assessment devices and practices that are in the public interest.

CG 014102

[1]Paper presented at the Symposium, Psychology and the Public Good, APA Convention, New York City, September 4, 1979.

The Interest Inventory Controversy

The interest inventory controversy began in the period 1971 to 1973, when individuals, groups, committees, and commissions charged that interest inventories served to keep women and men in traditional occupations. The belief was that interest inventories by virtue of their items, instructions, interpretative materials, scoring and normative procedures helped to maintain the sex-segregated character of the work force. This belief was reinforced by numerous plausible hypotheses and some plausible data. For example, when males and females of any age take interest inventories, the distributions of suggested occupations for males and females are usually somewhat divergent. According to interest inventories, a small proportion of females have skilled trades and factory-oriented interests, whereas men have those interests in larger proportions.

The first stage of this controversy was characterized by rational analyses by critics that inventories must exert a restrictive force on women and men, and by rational defenses by authors that these evil events were unlikely.

The second stage was characterized by recommendations and guidelines for reducing the alleged sex bias of inventories. This collection of ideas and recommendations of the National Institute of Education and the Association for Measurement and Evaluation in Guidance and later by Title IX and by the Office of Civil Rights appeared to chart some activities and test revisions that would change the alleged undesirable influences of interest inventories. Authors and publishers made multiple revisions that ranged from minor changes in items to major reconstructions of an entire inventory.

The third stage involved the empirical examination of the multiple recommendations and solutions that were proposed to reduce sex differences, to foster non-traditional goals, and to increase self-understanding of sex-role socialization. The need for empirical investigation became clear in

1972 when no one was able to locate a single experimental study of the influence of any interest inventory on the test-taker. The few existing evaluations were compounded by the effects of counselors and interest inventories as well as other tests or treatments. I will not review all of these more recent investigations (there are now about 50) but I will outline the main ideas and findings in the experimental evaluations.

These experiments usually entail the following general plan. Males and females are assessed for their current vocational aspirations, satisfaction with their choice, degree of decidedness, degree of self-understanding, or related ideas, then people are assigned randomly to different experimental treatments (take an interest inventory, see a professional counselor, use a vocational card sort, or receive no treatment), and finally after the treatment--either immediately or at intervals up to three months later--females and males are again polled for their vocational aspirations, ratings of the treatment, information-seeking activity and related criteria.

The initial group of experiments compared the effect of an interest inventory either with controls, or with some special treatment plus an interest inventory. These early experiments indicated that people who took interest inventories listed more vocational choices, increased their satisfaction with a current choice, and reported increased self-understanding. No sex differences were observed in any of these beneficial outcomes.

A second group of experiments has involved the comparison of different interest inventories (the Strong-Campbell Interest Inventory versus the Self Directed Search), vocational card sorts (you do your own categorizing and scoring), and group counseling treatments. Although these comparisons involve different inventories and treatments developed by divergent methods, and have different interpretative materials and scoring procedures, the results suggest that different vocational treatments have similar or identical effects on the test-taker or treatment taker. Again, no sex differences

have been observed in the beneficial outcomes.

A third group of experiments has involved piling several treatments on the client. For example, adding a vocational card sort to the SDS is no more effective than either device alone.

A fourth group of experiments, shich I characterize as "analytical", illuminate how the beneficial effects observed in the earlier shotgun investigations may occur and what changes may increase a person's options and satisfaction. A group of five simple experiments produced no significant mean differences in outcomes for males and females when the directions for an interest inventory were radically revised, and when all items were rendered gender neutral. Likewise, a sixth experiment indicated that the structure, self-scoring or sheer number of vocational options contained in an inventory may have no special merit.

These analytical investigations led to a recent experiment in which high school students were polled for their expectations for the interest inventory that they were about to take. Both females and males indicated that they wanted most of all, "reassurance" about an aspiration they already had. "Wanting more alternatives" ranked well below this desire. High school girls wanted more or fewer options to about the same degree. Boys wanted fewer options rather than more. In this same investigation, we found that students with a clear sense of identity and few vocational decision-making problems rated the helpfulness of the interest inventory higher than did students with a diffuse sense of identity and multiple decision-making difficulties.

Taken together, these more analytical experiments imply that the beneficial effects of inventories are due to the common elements present in divergent inventories, card sorts, and related treatments. These common elements include: (a) exposure to occupational information in items and interpretative materials, (b) cognitive rehearsal of vocational aspirations in the process of filling out inventories and talking to

counselors or others. And (c), the combination of occupational information and rehearsal probably stimulates conceptions of occupational structure and self-understanding. These events seem likely, because the average student knows very little about the occupational world so that a small amount of new information makes a big difference.

This ends my hurried review of the evidence about the experimental evaluations of interest inventories. (See Holland, 1979, for a more detailed review of these investigations). Parenthetically, if you are an aficionado of this controversy, you have noticed that I omitted any discussion of the raw score, normed score, and balanced scale proposals and disagreements. I have done so because the experimental comparisons of inventories that use divergent scale construction techniques have identical or similar effects. There has been no experimental test of the effect of any balanced scale inventory on the test-taker, but I expect the effects to be similar to those observed for normed inventories, because a balanced scale inventory, like a normed scale inventory, by virtue of its construction, presents similar distributions of options to males and females. And, the experimental comparisons of inventories that present similar or divergent distributions of vocational alternatives to females and males do not influence females or males in different ways. No significant interactions.

Now I want to emphasize three major findings in these experimental evaluations.

First, interest inventories have small but apparently beneficial effects on a wide range of populations (high school and college students, males and females, blacks and whites, Hispanics). There is no experimental evidence that documents a single negative effect. In addition, there is more experimental evidence to support the beneficial values of interest inventories than there is for counselors, vocational card sorts, or group counseling treatments.

Second, none of the ideas advocated in the NIE guidelines, AMEG, or Title IX regulations have resulted in inventory revisions that have a demonstrable effect on the user. In short, beliefs about the sex-fairor sex-bias characteristics of a particular inventory are not related to its impact on the test-taker. To the contrary, inventories with divergent origins and construction have similar, if not identical, effects.

Third, the distributions of vocational aspirations for females before and after taking an inventory are very similar. The same trend holds for males. In addition, these distributions for males and females are divergent both before and after taking an inventory. In short, inventories make only a slight dent, if at all, in the life long shaping forces that society exerts on females and males.

Some Implications

The political, regulatory, and research activity generated by this controversy appears to have some cogent implications for psychological testing generally. I have organized my interpretations of this experience according to the impact of testing on the public, related APA organizational and technical problems, and some common insights about testing practices.

Imapct of Testing

The controversy about interest inventories is a useful illustration of what we know and don't know about a particular kind of inventory. Despite the recent experimental evaluations, more is known about the reliability and construct validity of these inventories than their influence

on the public. To estimate the probable impact of interest inventories, it is necessary to string together the experimental studies of self reported impact and several ambiguous criteria such as the increasing popularity of these inventories from year to year and the absence of any legal action.

What is missing, is one or more national surveys that would permit generalizations like "70% of the people who take an interest inventory found it helpful, 20% didn't know, and 10% found the experience distressing or harmful." Surveys of this kind could be used to identify negative outcomes for review in more elaborate, clinical investigations. These investigations would provide reliable estimates of the parameters of satisfaction, benefit, and harmful effect for representative populations.

Our knowledge of the impact of intelligence, admissions, and employment tests which are now involved in intense controversy, has a similar pattern of strengths and weaknesses. More is known about these tests according to multiple investigations of reliability, validity and content than is known about their impact on the public.

The impact of intelligence, admissions, and employment tests must be estimated by piecing together statistical studies of probable impact of a test (if a test is used with a rigid adherence to a statistical formula), by a small number of legal rulings, and by anecdotal accounts of positive and negative outcomes.

This collection of evidence appears impossible to collate and evaluate. Again, some national surveys of test takers would provide a beginning for estimating the parameters of client satisfaction, strengths and weaknesses of tests and testing practices and for guiding a succession of more focused inquiries.

A more comprehensive and explicit knowledge of testing influences would stimulate the development of better tests and would provide information for an effective defense of sound tests and testing practices. In 1966, David Goslin outlined a comprehensive program of research for estimating

and comprehending the impact of testing. And in the last year, the National Academy of Sciences has established a Committee on Ability Testing that is investigating the use of aptitude tests in diverse settings. Things are moving, but often long after the controversy erupted.

Some APA Organizational and Technical Problems

The interest inventory controversy also revealed a number of organizational, political, and technical problems that require better solutions. And fortunately, the APA has begun to move on one of these difficulties--the APA Test Standards.

1. The current test standards have little to say about the need to assess the actual impact of tests on the test-taker. Instead, they stress the importance of reliability and validity. Fortunately, the Committee on tests has held hearings and is moving toward a new set of test standards.

2. The APA has budgeted 6,500 dollars this year for the Committee on Psychological Tests and Assessments. The Committee has six talented members who meet three times a year. This organizational commitment appears too small to provide intellectual leadership for an industry that administers perhaps 300,000,000 tests a year and may have annual budgets totaling 50 million.

3. During this 7-year controversy on interest measurement, the only public action that the APA took was to imply support for the NIE Guidelines by mailing them to members as guidelines to consider. This action gave the NIE Guidelines a scientific status they do not deserve. And, while the APGA, the NIE, and the AMEG commissions developed recommendations and regulations, our APA Committee appeared to be concerned with other matters and presumably overwhelmed by the multiple public test controversies.

Test publishers appeared to cope with this controversy by silence and revising inventories, but they soon discovered that revisions did not usually bring praise. Critics, authors, and publishers did not agree. Likewise, the brief paragraphs about testing materials in Title IX turned

out to be a new projective device--different people read different meanings into those terse paragraphs.

Common Insights

This controversy also made us more aware of the basic ingredients of beneficial test development and practice. It is still clear that a test's reliability and validity are vital characteristics, but it is also clear that the tested impact of a psychological device is equally important. In addition, the ethical and skilled use of tests requires an integration of values, assessment goals, instructional technology, and knowledge of the field in which the testing takes place. Testing has been dominated by psychometricians and statisticians. It seems more apparent now that psychologists need to be more concerned with the context of testing and take steps to think about tests and their role in the educational and social systems of information and use in which tests are imbedded. And, we need to integrate the experiences of test developers, psychometricians, practitioners, researchers and consumers.

There is a remarkable group of common denominators in the controversies about the effects of admissions, employment, intelligence, and interest tests. In each controversy, participants discoverd the ambiguities inherent in defining a particular kind of bias, the strengths and weaknesses of internal versus external criteria for estimating a test's impact, the role that values play in test construction and testing practices. Finally, these controversies have served to sensitize psychologists to "cultural, and racial differences and how apparently innocent and benign practices may perpetuate discrimination."

For instance, in 1973, many participants believed that the definition of sex-bias was a relatively simple explicit matter. Many thought that you can identify sex bias by simply looking at words and items, by seeing if the endorsement rates for items were about the same for females and males, by seeing if the distributions of interest inventory

outcomed for females and males were similar, or by conducting experiments to learn what happens to people when they experience inventories with different characteristics.

As these ideas diffused among the participants, other criteria for determining the presence or absence of bias were proposed: does the inventory in question increase vocational information-seeking? Selection of non-traditional options? Number and variety of occupations being considered? Satisfaction with a current vocational choice? Now, it is apparent that none of these and other outcomes are applicable to everyone or to people of all ages. Some people want more options; some want fewer options, and others want reassurance about an option they have in mind. Some people are unhappy with the outcomes of taking an interest inventory when it fails to support their choice. Sometimes, they should be--they want to do something that is foreign to their interests; other times the inventory is in error. A clinical review of the case data is needed and that review may end in ambiguity.

I could continue to elaborate the problems of defining bias, but I wanted only to illustrate some of the difficulties in defining bias, a problem that most participants came to appreciate but a problem that the Title IX and NIE Guidelines failed to cope with. These guidelines form vague nets to raise questions of bias, and they were written long before we had any explicit knowledge of the internal structure or effects of inventories.

One of the amusing incidents in this controversy (there were several) was that two advocates of precipitous and uninformed regulation fell into a secondary controversy about what Title IX really meant.

Some Recommendations

This seven-year controversy has stimulated some ideas for how we might cope with controversies about testing, and how we might gain a more accurate and comprehensive assessment of the impact of testing on the

public. I have already indicated some of these recommendations, but I would like to summarize them.

First, although we have some evidence about the actual impact of a few inventories, we need a comprehensive summary of the direct and indirect effects of taking tests of all kinds. Such an effort might entail a review of Goslin's earlier thinking and the planning of an extensive program of research. That kind of research should interest many federal and non-federal sponsors. The APA test committee might be charged with leading such an effort. Test publishers might be interested in contributing funds. Such research should be useful in estimating the strengths and weaknesses in tests and test practices that deserve praise and continuation, and those tests and testing practices that require revision.

Second, the APA Committee on Psychological Tests and Assessments should be enlarged and restructured to provide greater national leadership. The present committee of six is simply too small to deal with the multiple and increasing problems of testing. A committee with a reasonable chance of dealing with our current problems might include 30 members. The committee might be organized as six subcommittees: Legal Questions, Public Information, Aptitude and Achievement Tests, Personality and Interest Inventories, Controversy Task Force, and an Ex-officio Committee consisting of one large and one small test publisher, two disgruntled test authors, two bonafide practitioners who use tests but are not in awe of psychometricians, and two activists from the general public who do not own and operate a newspaper.

I do not believe such a committee is too large or too expensive, for as a recent APA resolution states, "Psychological testing represents one of our most valuable and practical contributions to society." Again, why not ask test publishers, testing agencies, authors, and federal agencies to contribute funds? The cost of a large committee could be reduced by rotating some subcommittees around the United States.

I have also daydreamed about some immediate tasks for these sub-

committees. For example, the public information committee might get out the good news. There is considerable good news to report, but it fails to get attention.

Here are two illustrative developments that have gone unreported by the press and that the public information committee could call attention to.

First, the Student Profile Section and the Student Characteristics section used in conjunction with the ACT and SAT admissions tests contain simple scales designed to assess a student's non academic potentials and originality and were developed to supplement the traditional criteria of academic test scores and grades. Also, Baird and Knapp (1979) at ETS have created a more elaborate form for the selection of graduate students. This form calls attention to a student's record of accomplishment outside the classroom. Taylor and his colleagues have developed materials for identifying six kinds of talents in classrooms; they have trained teachers how to stimulate the development of these talents (academic, creativity, planning, communicating, forecasting, and decision-making). And, they have performed multiple evaluations that demonstrate that these new dimensions, and the associated teaching training program have a positive effect on students. Put another way, some of the cogent criticisms of admissions and educational achievement tests--they fail to assess many important human potentials--have been dealt with for more than ten years, but these valuable developments have gone unnoticed.

Second, there are now at least four self-administered and self-scored interest inventories. That development has been going on for about eight years. The research implies that their reliability, validity and effects on the user are not better or worse than more conventional inventories that are scored by electronic machines. They represent a major attempt to make vocational assistance available to a larger proportion of the public.

The tasks for the remaining subcommittees are generally obvious, but

perhaps some elaboration of the role of the Task Force on Test Controversy is needed. This group of psychological firefighters could move to the site of intense controversies and act as fact-finders, teachers, referees, and conciliators. In principle, they would resemble the aircraft accident teams of the FAA. They might include a lawyer-psychologist, a clinician, a test expert, a social psychologist and a generalist. They might engage in fire prevention. For example, they might ask publishers to refrain from exploiting controversy. Currently, three publishers claim that their devices are sex-fair, because their inventories were developed in a special way, but they have no experimental evidence that these inventories affect females and males in the same way.

Third, the APA should take over the Buros Mental Measurement Yearbook. That Oscar Buros' contribution was magnificant, cannot be denied. At the same time, although these yearbooks provide much useful information, they also could benefit from a standard format for reviews and the elimination of some conflict of interest between reviewers and reviewees. Opinions and speeches should be clearly delineated, and publishers and authors should have the right to correct errors of fact. Many reviews and counter reviews simply confuse the potential user.

Other <u>less</u> satisfying Buros type yearbooks are in production. If the first imitation is a reliable sign of the future (sponsored by the Office of Education) we are in more trouble than ever. Their first effort relied largely on reliability and content validity.

I have only a one-sentence summary for my talk. I was told 13 pages or else. In a sentence, we must move from the era of reliability and construct validity to the era of reliability, validity <u>and</u> public impact.

References

Educational Testing Service. Inventory of activities and accomplishments. Princeton, New Jersey: Educational Testing Service, 1978.

Goslin, D. A. The Search for Ability. Wiley, 1966.

Holland, J. L. Professional Manual for the Self Directed Search (revised edition) Palo Alto, California: Consulting Psychologists Press, 1979.

Taylor, C. W. Teaching for talents and gifts. Salt Lake City: Utah State Board of Education, 1978.